1. STATISTICS

Revised September 2009 by G. Cowan (RHUL).

There are two main approaches to statistical inference, which we may call frequentist and Bayesian. In frequentist statistics, probability is interpreted as the frequency of the outcome of a repeatable experiment. The most important tools in this framework are parameter estimation, covered in Section 36.1, and statistical tests, discussed in Section 36.2. Frequentist confidence intervals, which are constructed so as to cover the true value of a parameter with a specified probability, are treated in Section 36.3.2. Note that in frequentist statistics one does not define a probability for a hypothesis or for a parameter.

In Bayesian statistics, the interpretation of probability is more general and includes *degree of belief* (called subjective probability). One can then speak of a probability density function (p.d.f.) for a parameter, which expresses one's state of knowledge about where its true value lies. Using Bayes' theorem Eq. (31.4), the prior degree of belief is updated by the data from the experiment. Bayesian methods for interval estimation are discussed in Sections 36.3.1 and 36.3.2.6

Following common usage in physics, the word "error" is often used in this chapter to mean "uncertainty." More specifically it can indicate the size of an interval as in "the standard error" or "error propagation," where the term refers to the standard deviation of an estimator.

1.1. Parameter estimation

Here we review the frequentist approach to *point estimation*of parameters. An *estimator* $\widehat{\theta}$ (written with a hat) is a function of the data whose value, the *estimate*, is intended as a meaningful guess for the value of the parameter θ.

1.1.1. ***Estimators for mean, variance and median***: Suppose we have a set of N independent measurements, x_i, assumed to be unbiased measurements of the same unknown quantity μ with a common, but unknown, variance σ^2. Then

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.4}$$

$$\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \widehat{\mu})^2 \tag{1.5}$$

are unbiased estimators of μ and σ^2. The variance of $\widehat{\mu}$ is σ^2/N and the variance of $\widehat{\sigma^2}$ is

$$V\left[\widehat{\sigma^2}\right] = \frac{1}{N}\left(m_4 - \frac{N-3}{N-1}\sigma^4\right) , \tag{1.6}$$

where m_4 is the 4th central moment of x. For Gaussian distributed x_i, this becomes $2\sigma^4/(N-1)$ for any $N \geq 2$, and for large N, the standard deviation of $\widehat{\sigma}$ (the "error of the error") is $\sigma/\sqrt{2N}$. Again, if the x_i are Gaussian, $\widehat{\mu}$ is an efficient estimator for μ, and the estimators $\widehat{\mu}$ and $\widehat{\sigma^2}$ are uncorrelated. Otherwise the arithmetic mean (1.4) is not necessarily the most efficient estimator.

If the x_i have different, known variances σ_i^2, then the weighted average

$$\widehat{\mu} = \frac{1}{w}\sum_{i=1}^{N} w_i x_i \tag{1.7}$$

is an unbiased estimator for μ with a smaller variance than an unweighted average; here $w_i = 1/\sigma_i^2$ and $w = \sum_i w_i$. The standard deviation of $\widehat{\mu}$ is $1/\sqrt{w}$.

1.1.2. ***The method of maximum likelihood***: Suppose we have a set of N measured quantities $\boldsymbol{x} = (x_1, \ldots, x_N)$ described by a joint p.d.f. $f(\boldsymbol{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is set of n parameters whose values are unknown. The *likelihood function* is given by the p.d.f. evaluated with the data $\boldsymbol{x}$, but viewed as a function of the parameters, *i.e.*, $L(\boldsymbol{\theta}) = f(\boldsymbol{x}; \boldsymbol{\theta})$. If the measurements x_i are statistically independent and each follow the p.d.f. $f(x; \boldsymbol{\theta})$, then the joint p.d.f. for $\boldsymbol{x}$ factorizes and the likelihood function is

$$L\left(\boldsymbol{\theta}\right) = \prod_{i=1}^{N} f\left(x_i; \boldsymbol{\theta}\right) \ . \tag{1.8}$$

The method of maximum likelihood takes the estimators $\widehat{\boldsymbol{\theta}}$ to be those values of $\boldsymbol{\theta}$ that maximize $L(\boldsymbol{\theta})$.

Note that the likelihood function is *not* a p.d.f. for the parameters $\boldsymbol{\theta}$; in frequentist statistics this is not defined. In Bayesian statistics, one can obtain from the likelihood the posterior p.d.f. for $\boldsymbol{\theta}$, but this requires multiplying by a prior p.d.f. (see Sec. 36.3.1).

It is usually easier to work with $\ln L$, and since both are maximized for the same parameter values $\boldsymbol{\theta}$, the maximum likelihood (ML) estimators can be found by solving the *likelihood equations*,

$$\frac{\partial \ln L}{\partial \theta_i} = 0 \ , \qquad i = 1, \ldots, n \ . \tag{1.9}$$

In evaluating the likelihood function, it is important that any normalization factors in the p.d.f. that involve $\boldsymbol{\theta}$ be included.

The inverse V^{-1} of the covariance matrix $V_{ij} = \mathrm{cov}[\widehat{\theta}_i, \widehat{\theta}_j]$ for a set of ML estimators can be estimated by using

$$\left(\widehat{V}^{-1}\right)_{ij} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\widehat{\theta}} \ . \tag{1.10}$$

For finite samples, however, Eq. (1.10) can result in an underestimate of the variances. In the large sample limit (or in a linear model with Gaussian errors), L has a Gaussian form and $\ln L$ is (hyper)parabolic. In this case, it can be seen that a numerically equivalent way of determining s-standard-deviation errors is from the contour given by the $\boldsymbol{\theta}'$ such that

$$\ln L\left(\boldsymbol{\theta}'\right) = \ln L_{\max} - s^2/2 \ , \tag{1.11}$$

where $\ln L_{\max}$ is the value of $\ln L$ at the solution point (compare with Eq. (36.58)). The extreme limits of this contour on the θ_i axis give an approximate s-standard-deviation confidence interval for θ_i (see Section 36.3.2.4).

1.1.3. ***The method of least squares***: The *method of least squares* (LS) coincides with the method of maximum likelihood in the following special case. Consider a set of N independent measurements y_i at known points x_i. The measurement y_i is assumed to be Gaussian distributed with mean $F(x_i; \boldsymbol{\theta})$ and known variance σ_i^2. The goal is

to construct estimators for the unknown parameters $\boldsymbol{\theta}$. The likelihood function contains the sum of squares

$$\chi^2(\boldsymbol{\theta}) = -2\ln L(\boldsymbol{\theta}) + \text{ constant } = \sum_{i=1}^{N} \frac{(y_i - F(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2} . \quad (1.13)$$

The set of parameters $\boldsymbol{\theta}$ which maximize L is the same as those which minimize χ^2.

The minimum of Equation (1.13) defines the least-squares estimators $\widehat{\boldsymbol{\theta}}$ for the more general case where the y_i are not Gaussian distributed as long as they are independent. If they are not independent but rather have a covariance matrix $V_{ij} = \text{cov}[y_i, y_j]$, then the LS estimators are determined by the minimum of

$$\chi^2(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{F}(\boldsymbol{\theta}))^T V^{-1} (\boldsymbol{y} - \boldsymbol{F}(\boldsymbol{\theta})) , \quad (1.14)$$

where $\boldsymbol{y} = (y_1, \ldots, y_N)$ is the vector of measurements, $\boldsymbol{F}(\boldsymbol{\theta})$ is the corresponding vector of predicted values (understood as a column vector in (1.14)), and the superscript T denotes transposed (*i.e.*, row) vector.

In many practical cases, one further restricts the problem to the situation where $F(x_i;\boldsymbol{\theta})$ is a linear function of the parameters, *i.e.*,

$$F(x_i;\boldsymbol{\theta}) = \sum_{j=1}^{m} \theta_j h_j(x_i) . \quad (1.15)$$

Here the $h_j(x)$ are m linearly independent functions, *e.g.*, $1, x, x^2, \ldots, x^{m-1}$, or Legendre polynomials. We require $m < N$ and at least m of the x_i must be distinct.

Minimizing χ^2 in this case with m parameters reduces to solving a system of m linear equations. Defining $H_{ij} = h_j(x_i)$ and minimizing χ^2 by setting its derivatives with respect to the θ_i equal to zero gives the LS estimators,

$$\widehat{\boldsymbol{\theta}} = \left(H^T V^{-1} H\right)^{-1} H^T V^{-1} \boldsymbol{y} \equiv D\boldsymbol{y} . \quad (1.16)$$

The covariance matrix for the estimators $U_{ij} = \text{cov}[\widehat{\theta}_i, \widehat{\theta}_j]$ is given by

$$U = DVD^T = \left(H^T V^{-1} H\right)^{-1} . \quad (1.17)$$

Expanding $\chi^2(\boldsymbol{\theta})$ about $\widehat{\boldsymbol{\theta}}$, one finds that the contour in parameter space defined by

$$\chi^2(\boldsymbol{\theta}) = \chi^2\left(\widehat{\boldsymbol{\theta}}\right) + 1 = \chi^2_{\text{min}} + 1 \quad (1.23)$$

has tangent planes located at approximately plus-or-minus-one standard deviation $\sigma_{\widehat{\theta}}$ from the LS estimates $\widehat{\boldsymbol{\theta}}$.

As the minimum value of the χ^2 represents the level of agreement between the measurements and the fitted function, it can be used for assessing the goodness-of-fit; this is discussed further in Section 36.2.2.

1.1.5. ***Propagation of errors***: Consider a set of n quantities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and a set of m functions $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_m(\boldsymbol{\theta}))$. Suppose we have estimated $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$, using, say, maximum-likelihood or least-squares, and we also know or have estimated the covariance matrix $V_{ij} = \text{cov}[\widehat{\theta}_i, \widehat{\theta}_j]$. The goal of *error propagation* is to determine the covariance matrix for the functions, $U_{ij} = \text{cov}[\widehat{\eta}_i, \widehat{\eta}_j]$, where $\widehat{\boldsymbol{\eta}} = \boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})$. In particular, the diagonal elements $U_{ii} = V[\widehat{\eta}_i]$ give the variances. The new covariance matrix can be found by expanding the functions $\boldsymbol{\eta}(\boldsymbol{\theta})$ about the estimates $\widehat{\boldsymbol{\theta}}$ to first order in a Taylor

series. Using this one finds

$$U_{ij} \approx \sum_{k,l} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \bigg|_{\widehat{\theta}} V_{kl} \; . \tag{1.29}$$

This can be written in matrix notation as $U \approx AVA^T$ where the matrix of derivatives A is

$$A_{ij} = \frac{\partial \eta_i}{\partial \theta_j} \bigg|_{\widehat{\theta}} \; , \tag{1.30}$$

and A^T is its transpose. The approximation is exact if $\boldsymbol{\eta}(\boldsymbol{\theta})$ is linear.

1.2. Statistical tests

1.2.1. ***Hypothesis tests***: Consider an experiment whose outcome is characterized by a vector of data $\boldsymbol{x}$. A *hypothesis* is a statement about the distribution of $\boldsymbol{x}$. It could, for example, define completely the p.d.f. for the data (a simple hypothesis), or it could specify only the functional form of the p.d.f., with the values of one or more parameters left open (a composite hypothesis).

A *statistical test* is a rule that states for which values of $\boldsymbol{x}$ a given hypothesis (often called the null hypothesis, H_0) should be rejected in favor of its alternative H_1. This is done by defining a region of $\boldsymbol{x}$-space called the critical region; if the outcome of the experiment lands in this region, H_0 is rejected, otherwise it is accepted.

Rejecting H_0 if it is true is called an error of the first kind. The probability for this to occur is called the *size* or *significance level* of the test, α, which is chosen to be equal to some pre-specified value. It can also happen that H_0 is false and the true hypothesis is the alternative, H_1. If H_0 is accepted in such a case, this is called an error of the second kind, which will have some probability β. The quantity $1 - \beta$ is called the *power* of the test relative to H_1.

Often one tries to construct a test to maximize power for a given significance level, *i.e.*, to maximize the signal efficiency for a given significance level. The *Neyman–Pearson lemma* states that this is done by defining the acceptance region such that, for $\boldsymbol{x}$ in that region, the ratio of p.d.f.s for the hypotheses H_1 (signal) and H_0 (background),

$$\lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)} \; , \tag{1.31}$$

is greater than a given constant, the value of which is chosen to give the desired signal efficiency. Here H_0 and H_1 must be simple hypotheses, i.e., they should not contain undetermined parameters. The lemma is equivalent to the statement that (1.31) represents the test statistic with which one may obtain the highest signal efficiency for a given purity for the selected sample. It can be difficult in practice, however, to determine $\lambda(\boldsymbol{x})$, since this requires knowledge of the joint p.d.f.s $f(\boldsymbol{x}|H_0)$ and $f(\boldsymbol{x}|H_1)$.

In the usual case where the likelihood ratio (1.31) cannot be used explicitly, there exist a variety of other multivariate classifiers that effectively separate different types of events. Methods often used in HEP include *neural networks* or *Fisher discriminants* (see [10]). Recently, further classification methods from machine-learning have been applied in HEP analyses; these include *probability density estimation (PDE)* techniques, *kernel-based PDE* (*KDE* or *Parzen window*), *support vector machines*, and *decision trees*. Techniques such as "boosting" and "bagging" can be applied to combine a number of

classifiers into a stronger one with greater stability with respect to fluctuations in the training data.

1.2.2. ***Significance tests***: Often one wants to quantify the level of agreement between the data and a hypothesis without explicit reference to alternative hypotheses. This can be done by defining a statistic t, which is a function of the data whose value reflects in some way the level of agreement between the data and the hypothesis.

The hypothesis in question, say, H_0, will determine the p.d.f. $g(t|H_0)$ for the statistic. The significance of a discrepancy between the data and what one expects under the assumption of H_0 is quantified by giving the p-value, defined as the probability to find t in the region of equal or lesser compatibility with H_0 than the level of compatibility observed with the actual data. For example, if t is defined such that large values correspond to poor agreement with the hypothesis, then the p-value would be

$$p = \int_{t_{\rm obs}}^{\infty} g\left(t|H_0\right)\, dt \ , \tag{1.32}$$

where $t_{\rm obs}$ is the value of the statistic obtained in the actual experiment. The p-value should not be confused with the size (significance level) of a test, or the confidence level of a confidence interval (Section 36.3), both of which are pre-specified constants.

The p-value is a function of the data, and is therefore itself a random variable. If the hypothesis used to compute the p-value is true, then for continuous data, p will be uniformly distributed between zero and one. Note that the p-value is not the probability for the hypothesis; in frequentist statistics, this is not defined. Rather, the p-value is the probability, under the assumption of a hypothesis H_0, of obtaining data at least as incompatible with H_0 as the data actually observed.

When estimating parameters using the method of least squares, one obtains the minimum value of the quantity χ^2 (1.13). This statistic can be used to test the *goodness-of-fit*, *i.e.*, the test provides a measure of the significance of a discrepancy between the data and the hypothesized functional form used in the fit. It may also happen that no parameters are estimated from the data, but that one simply wants to compare a histogram, *e.g.*, a vector of Poisson distributed numbers $\boldsymbol{n} = (n_1, \ldots, n_N)$, with a hypothesis for their expectation values $\nu_i = E[n_i]$. As the distribution is Poisson with variances $\sigma_i^2 = \nu_i$, the χ^2 (1.13) becomes *Pearson's* χ^2 *statistic*,

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} \ . \tag{1.34}$$

If the hypothesis $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$ is correct, and if the expected values ν_i in (1.34) are sufficiently large (in practice, this will be a good approximation if all $\nu_i > 5$), then the χ^2 statistic will follow the χ^2 p.d.f. with the number of degrees of freedom equal to the number of measurements N minus the number of fitted parameters. The minimized χ^2 from Eq. (1.13) also has this property if the measurements y_i are Gaussian.

Assuming the goodness-of-fit statistic follows a χ^2 p.d.f., the p-value for the hypothesis is then

$$p = \int_{\chi^2}^{\infty} f\left(z; n_{\rm d}\right)\, dz \ , \tag{1.35}$$

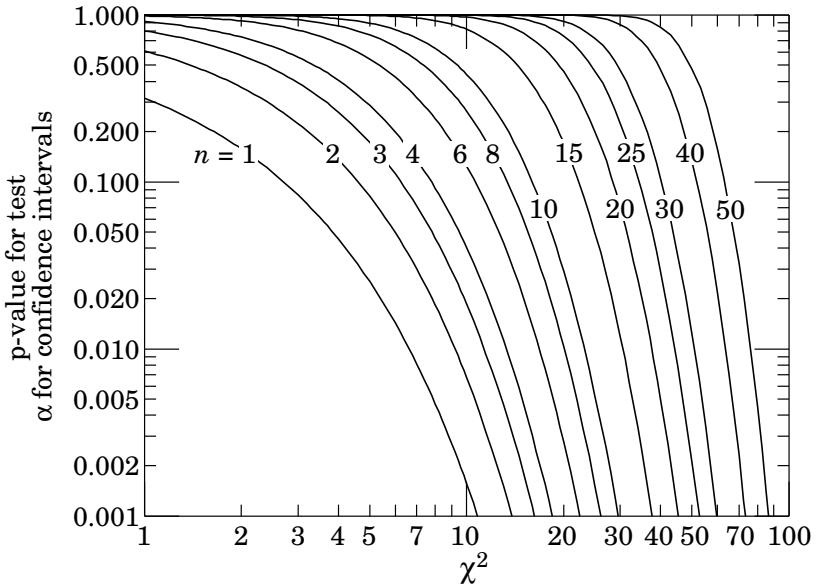

Figure 1.1: One minus the χ^2 cumulative distribution, $1 - F(\chi^2; n)$, for n degrees of freedom. This gives the p-value for the χ^2 goodness-of-fit test as well as one minus the coverage probability for confidence regions (see Sec. 36.3.2.4).

where $f(z; n_\mathrm{d})$ is the χ^2 p.d.f. and n_d is the appropriate number of degrees of freedom. Values can be obtained from Fig. 36.1 or from the CERNLIB routine `PROB` or the ROOT function `TMath::Prob`.

Since the mean of the χ^2 distribution is equal to n_d, one expects in a "reasonable" experiment to obtain $\chi^2 \approx n_\mathrm{d}$. Hence the quantity χ^2/n_d is sometimes reported. Since the p.d.f. of χ^2/n_d depends on n_d, however, one must report n_d as well if one wishes to determine the p-value. The p-values obtained for different values of χ^2/n_d are shown in Fig. 36.2.

1.2.3. ***Bayesian model selection***: In Bayesian statistics, all of one's knowledge about a model is contained in its posterior probability, which one obtains using Bayes' theorem. Thus one could reject a hypothesis H if its posterior probability $P(H|\boldsymbol{x})$ is sufficiently small. The difficulty here is that $P(H|\boldsymbol{x})$ is proportional to the prior probability $P(H)$, and there will not be a consensus about the prior probabilities for the existence of new phenomena. Nevertheless one can construct a quantity called the Bayes factor (described below), which can be used to quantify the degree to which the data prefer one hypothesis over another, and is independent of their prior probabilities.

Consider two models (hypotheses), H_i and H_j, described by vectors of parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, respectively. Some of the components will be common to both models and others may be distinct. The full prior probability for each model can be written in the form

$$\pi\left(H_i, \boldsymbol{\theta}_i\right) = P\left(H_i\right) \pi\left(\boldsymbol{\theta}_i | H_i\right) \ , \tag{1.36}$$

Here $P(H_i)$ is the overall prior probability for H_i, and $\pi(\boldsymbol{\theta}_i|H_i)$ is the normalized p.d.f. of its parameters. For each model, the posterior probability is found using Bayes' theorem,

$$P\left(H_i|\boldsymbol{x}\right) = \frac{\int L\left(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i\right) P\left(H_i\right) \pi\left(\boldsymbol{\theta}_i|H_i\right) \, d\boldsymbol{\theta}_i}{P\left(\boldsymbol{x}\right)} \ , \tag{1.37}$$

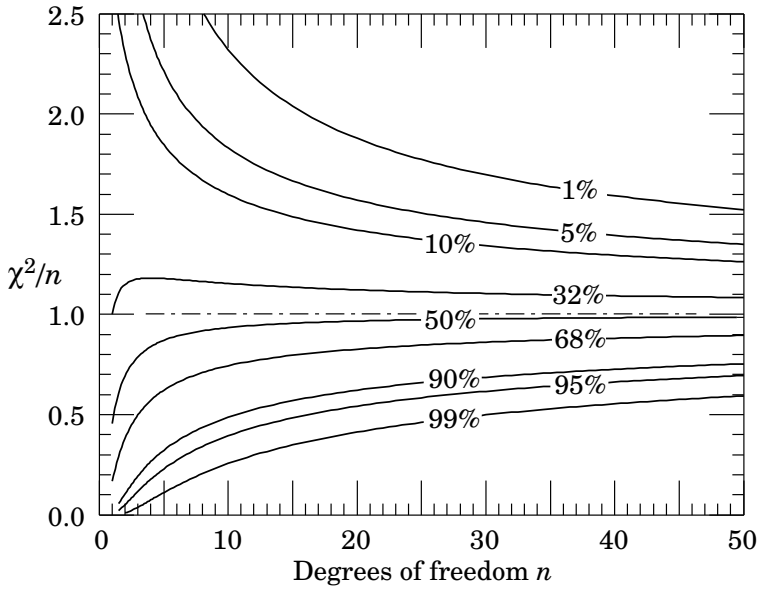

Figure 1.2: The 'reduced' χ^2, equal to χ^2/n, for n degrees of freedom. The curves show as a function of n the χ^2/n that corresponds to a given p-value.

where the integration is carried out over the internal parameters $\boldsymbol{\theta}_i$ of the model. The ratio of posterior probabilities for the models is therefore

$$\frac{P\left(H_i|\boldsymbol{x}\right)}{P\left(H_j|\boldsymbol{x}\right)} = \frac{\int L\left(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i\right) \pi\left(\boldsymbol{\theta}_i|H_i\right)\, d\boldsymbol{\theta}_i}{\int L\left(\boldsymbol{x}|\boldsymbol{\theta}_j, H_j\right) \pi\left(\boldsymbol{\theta}_j|H_j\right)\, d\boldsymbol{\theta}_j} \frac{P\left(H_i\right)}{P\left(H_j\right)} \,. \tag{1.38}$$

The *Bayes factor* is defined as

$$B_{ij} = \frac{\int L\left(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i\right) \pi\left(\boldsymbol{\theta}_i|H_i\right)\, d\boldsymbol{\theta}_i}{\int L\left(\boldsymbol{x}|\boldsymbol{\theta}_j, H_j\right) \pi\left(\boldsymbol{\theta}_j|H_j\right)\, d\boldsymbol{\theta}_j} \,. \tag{1.39}$$

This gives what the ratio of posterior probabilities for models i and j would be if the overall prior probabilities for the two models were equal. If the models have no nuisance parameters *i.e.*, no internal parameters described by priors, then the Bayes factor is simply the likelihood ratio. The Bayes factor therefore shows by how much the probability ratio of model i to model j changes in the light of the data, and thus can be viewed as a numerical measure of evidence supplied by the data in favour of one hypothesis over the other.

Although the Bayes factor is by construction independent of the overall prior probabilities $P(H_i)$ and $P(H_j)$, it does require priors for all internal parameters of a model, *i.e.*, one needs the functions $\pi(\boldsymbol{\theta}_i|H_i)$ and $\pi(\boldsymbol{\theta}_j|H_j)$. In a Bayesian analysis where one is only interested in the posterior p.d.f. of a parameter, it may be acceptable to take an unnormalizable function for the prior (an improper prior) as long as the product of likelihood and prior can be normalized. But improper priors are only defined up to an arbitrary multiplicative constant, which does not cancel in the ratio (1.39). Furthermore, although the range of a constant normalized prior is unimportant for parameter determination (provided it is wider than the likelihood), this is not so for the Bayes factor when such a prior is used for only one of the hypotheses. So to compute a Bayes factor, all internal parameters must be described by normalized priors that represent meaningful probabilities over the entire range where they are defined.

An exception to this rule may be considered when the identical parameter appears in the models for both numerator and denominator of the Bayes factor. In this case one can argue that the arbitrary constants would cancel. One must exercise some caution, however, as parameters with the same name and physical meaning may still play different roles in the two models. Both integrals in equation (1.39) are of the form

$$m = \int L\left(\boldsymbol{x}|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right)\, d\boldsymbol{\theta}\ , \tag{1.40}$$

which is called the *marginal likelihood* (or in some fields called the *evidence*). A review of Bayes factors including a discussion of computational issues is Ref. [30].

1.3. Intervals and limits

When the goal of an experiment is to determine a parameter θ, the result is usually expressed by quoting, in addition to the point estimate, some sort of interval which reflects the statistical precision of the measurement. In the simplest case, this can be given by the parameter's estimated value $\widehat{\theta}$ plus or minus an estimate of the standard deviation of $\widehat{\theta}$, $\sigma_{\widehat{\theta}}$. If, however, the p.d.f. of the estimator is not Gaussian or if there are physical boundaries on the possible values of the parameter, then one usually quotes instead an interval according to one of the procedures described below.

1.3.1. ***Bayesian intervals***: As described in Sec. 36.1.4, a Bayesian posterior probability may be used to determine regions that will have a given probability of containing the true value of a parameter. In the single parameter case, for example, an interval (called a Bayesian or credible interval) $[\theta_{\rm lo}, \theta_{\rm up}]$ can be determined which contains a given fraction $1-\alpha$ of the posterior probability, *i.e.*,

$$1-\alpha = \int_{\theta_{\rm lo}}^{\theta_{\rm up}} p\left(\theta|\boldsymbol{x}\right)\, d\theta\ . \tag{1.41}$$

Sometimes an upper or lower limit is desired, *i.e.*, $\theta_{\rm lo}$ can be set to zero or $\theta_{\rm up}$ to infinity. In other cases, one might choose $\theta_{\rm lo}$ and $\theta_{\rm up}$ such that $p(\theta|\boldsymbol{x})$ is higher everywhere inside the interval than outside; these are called *highest posterior density* (HPD) intervals. Note that HPD intervals are not invariant under a nonlinear transformation of the parameter.

If a parameter is constrained to be non-negative, then the prior p.d.f. can simply be set to zero for negative values. An important example is the case of a Poisson variable n, which counts signal events with unknown mean s, as well as background with mean b, assumed known. For the signal mean s, one often uses the prior

$$\pi\left(s\right) = \begin{cases} 0 & s < 0 \\ 1 & s \geq 0 \end{cases}\ . \tag{1.42}$$

In the absence of a clear discovery, (*e.g.*, if $n = 0$ or if in any case n is compatible with the expected background), one usually wishes to place an upper limit on s (see, however, Sec. 36.3.2.6 on "flip-flopping" concerning frequentist coverage). Using the likelihood function for Poisson distributed n,

$$L\left(n|s\right) = \frac{(s+b)^n}{n!} e^{-(s+b)}\ , \tag{1.43}$$

along with the prior (1.42) in (36.24) gives the posterior density for s. An upper limit $s_{\rm up}$ at confidence level (or here, rather, credibility

level) $1-\alpha$ can be obtained by requiring

$$1-\alpha = \int_{-\infty}^{s_{\rm up}} p\left(s|n\right) ds = \frac{\int_{-\infty}^{s_{\rm up}} L\left(n|s\right)\, \pi\left(s\right)\, ds}{\int_{-\infty}^{\infty} L\left(n|s\right)\, \pi\left(s\right)\, ds} \,, \tag{1.44}$$

where the lower limit of integration is effectively zero because of the cut-off in $\pi(s)$. By relating the integrals in Eq. (1.44) to incomplete gamma functions, the equation reduces to

$$\alpha = e^{-s_{\rm up}} \frac{\sum_{m=0}^{n} \left(s_{\rm up}+b\right)^m /m!}{\sum_{m=0}^{n} b^m/m!} \,. \tag{1.45}$$

This must be solved numerically for the limit $s_{\rm up}$. For the special case of $b=0$, the sums can be related to the *quantile* $F^{-1}_{\chi^2}$ of the χ^2 distribution (inverse of the cumulative distribution) to give

$$s_{\rm up} = \tfrac{1}{2} F^{-1}_{\chi^2}\left(1-\alpha; n_{\rm d}\right) \,, \tag{1.46}$$

where the number of degrees of freedom is $n_{\rm d} = 2(n+1)$. The quantile of the χ^2 distribution can be obtained using the CERNLIB routine `CHISIN`, or the ROOT function `TMath::ChisquareQuantile`. It so happens that for the case of $b=0$, the upper limits from Eq. (1.46) coincide numerically with the values of the frequentist upper limits discussed in Section 36.3.2.5. Values for $1-\alpha = 0.9$ and 0.95 are given by the values $\nu_{\rm up}$ in Table 36.3.

As in any Bayesian analysis, it is important to show how the result would change if one uses different prior probabilities. For example, one could consider the Jeffreys prior as described in Sec. 36.1.4. For this problem one finds the Jeffreys prior $\pi(s) \propto 1/\sqrt{s+b}$ for $s \geq 0$ and zero otherwise. As with the constant prior, one would not regard this as representing one's prior beliefs about s, both because it is improper and also as it depends on b. Rather it is used with Bayes' theorem to produce an interval whose frequentist properties can be studied.

1.3.2. *Frequentist confidence intervals*:

1.3.2.1. *The Neyman construction for confidence intervals:* Consider a p.d.f. $f(x;\theta)$ where x represents the outcome of the experiment and θ is the unknown parameter for which we want to construct a confidence interval. The variable x could (and often does) represent an estimator for θ. Using $f(x;\theta)$, we can find for a pre-specified probability $1-\alpha$, and for every value of θ, a set of values $x_1(\theta,\alpha)$ and $x_2(\theta,\alpha)$ such that

$$P\left(x_1 < x < x_2; \theta\right) = 1-\alpha = \int_{x_1}^{x_2} f\left(x;\theta\right)\, dx \,. \tag{1.47}$$

This is illustrated in Fig. 36.3: a horizontal line segment $[x_1(\theta,\alpha)$, $x_2(\theta,\alpha)]$ is drawn for representative values of θ. The union of such intervals for all values of θ, designated in the figure as $D(\alpha)$, is known as the *confidence belt*. Typically the curves $x_1(\theta,\alpha)$ and $x_2(\theta,\alpha)$ are monotonic functions of θ, which we assume for this discussion.

Upon performing an experiment to measure x and obtaining a value x_0, one draws a vertical line through x_0. The confidence interval for θ is the set of all values of θ for which the corresponding line segment $[x_1(\theta,\alpha), x_2(\theta,\alpha)]$ is intercepted by this vertical line. Such confidence intervals are said to have a *confidence level* (CL) equal to $1-\alpha$.

Now suppose that the true value of θ is θ_0, indicated in the figure. We see from the figure that θ_0 lies between $\theta_1(x)$ and $\theta_2(x)$ if and only if x lies between $x_1(\theta_0)$ and $x_2(\theta_0)$. The two events thus have the same probability, and since this is true for any value θ_0, we can

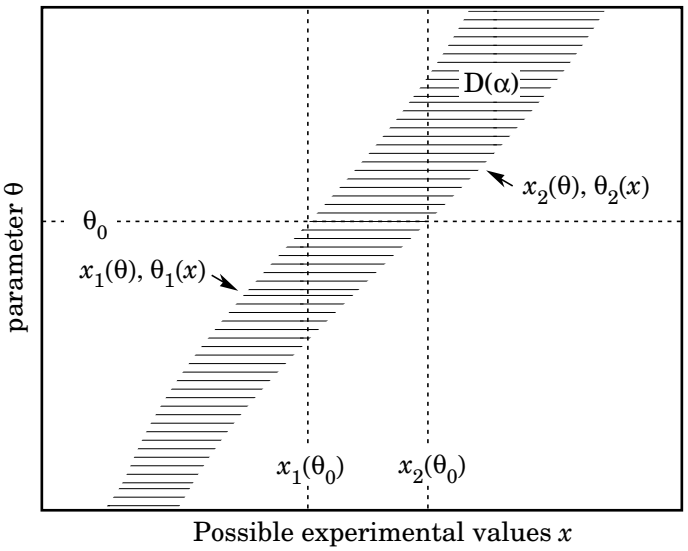

Figure 1.3: Construction of the confidence belt (see text).

drop the subscript 0 and obtain

$$1-\alpha = P\left(x_1\left(\theta\right) < x < x_2\left(\theta\right)\right) = P\left(\theta_2\left(x\right) < \theta < \theta_1\left(x\right)\right) \ . \quad (1.48)$$

In this probability statement, $\theta_1(x)$ and $\theta_2(x)$, *i.e.*, the endpoints of the interval, are the random variables and θ is an unknown constant. If the experiment were to be repeated a large number of times, the interval $[\theta_1, \theta_2]$ would vary, covering the fixed value θ in a fraction $1-\alpha$ of the experiments.

The condition of coverage in Eq. (1.47) does not determine x_1 and x_2 uniquely, and additional criteria are needed. The most common criterion is to choose *central intervals* such that the probabilities excluded below x_1 and above x_2 are each $\alpha/2$. In other cases, one may want to report only an upper or lower limit, in which case the probability excluded below x_1 or above x_2 can be set to zero. Another principle based on *likelihood ratio ordering* for determining which values of x should be included in the confidence belt is discussed in Sec. 1.3.2.2

When the observed random variable x is continuous, the coverage probability obtained with the Neyman construction is $1-\alpha$, regardless of the true value of the parameter. If x is discrete, however, it is not possible to find segments $[x_1(\theta,\alpha), x_2(\theta,\alpha)]$ that satisfy Eq. (1.47) exactly for all values of θ. By convention, one constructs the confidence belt requiring the probability $P(x_1 < x < x_2)$ to be *greater than or equal to* $1-\alpha$. This gives confidence intervals that include the true parameter with a probability greater than or equal to $1-\alpha$.

1.3.2.4. *Gaussian distributed measurements:* An important example of constructing a confidence interval is when the data consists of a single random variable x that follows a Gaussian distribution; this is often the case when x represents an estimator for a parameter and one has a sufficiently large data sample. If there is more than one parameter being estimated, the multivariate Gaussian is used. For the univariate case with known σ,

$$1-\alpha = \frac{1}{\sqrt{2\pi}\sigma}\int_{\mu-\delta}^{\mu+\delta} e^{-(x-\mu)^2/2\sigma^2}\,dx = \mathrm{erf}\left(\frac{\delta}{\sqrt{2}\,\sigma}\right) \quad (1.53)$$

is the probability that the measured value x will fall within $\pm\delta$ of the true value μ. From the symmetry of the Gaussian with respect to x and μ, this is also the probability for the interval $x \pm \delta$ to include μ. Fig. 36.4 shows a $\delta = 1.64\sigma$ confidence interval unshaded. The choice $\delta = \sigma$ gives an interval called the *standard error* which has $1 - \alpha = 68.27\%$ if σ is known. Values of α for other frequently used choices of δ are given in Table 36.1.

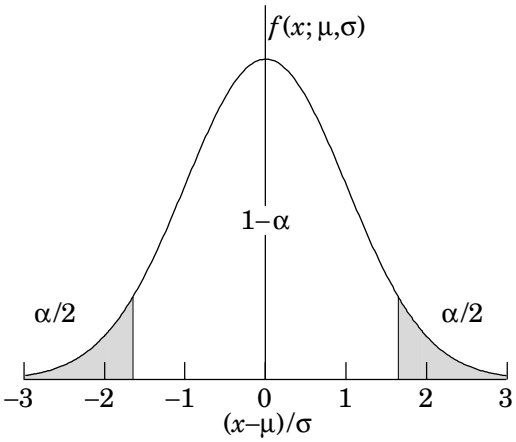

Figure 1.4: Illustration of a symmetric 90% confidence interval (unshaded) for a measurement of a single quantity with Gaussian errors. Integrated probabilities, defined by α, are as shown.

Table 1.1: Area of the tails α outside $\pm\delta$ from the mean of a Gaussian distribution.

α	δ	α	δ
0.3173	1σ	0.2	1.28σ
4.55×10^{-2}	2σ	0.1	1.64σ
2.7×10^{-3}	3σ	0.05	1.96σ
6.3×10^{-5}	4σ	0.01	2.58σ
5.7×10^{-7}	5σ	0.001	3.29σ
2.0×10^{-9}	6σ	10^{-4}	3.89σ

We can set a one-sided (upper or lower) limit by excluding above $x + \delta$ (or below $x - \delta$). The values of α for such limits are half the values in Table 1.1.

The relation (1.53) can be re-expressed using the cumulative distribution function for the χ^2 distribution as

$$\alpha = 1 - F\left(\chi^2; n\right) , \tag{1.54}$$

for $\chi^2 = (\delta/\sigma)^2$ and $n = 1$ degree of freedom. This can be obtained from Fig. 1.1 on the $n = 1$ curve or by using the CERNLIB routine `PROB` or the ROOT function `TMath::Prob`.

For multivariate measurements of, say, n parameter estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$, one requires the full covariance matrix $V_{ij} = \mathrm{cov}[\widehat{\theta}_i, \widehat{\theta}_j]$, which can be estimated as described in Sections 1.1.2 and 1.1.3. Under

fairly general conditions with the methods of maximum-likelihood or least-squares in the large sample limit, the estimators will be distributed according to a multivariate Gaussian centered about the true (unknown) values $\boldsymbol{\theta}$, and furthermore, the likelihood function itself takes on a Gaussian shape.

The standard error ellipse for the pair $(\widehat{\theta}_i, \widehat{\theta}_j)$ is shown in Fig. 36.5, corresponding to a contour $\chi^2 = \chi^2_{\rm min} + 1$ or $\ln L = \ln L_{\rm max} - 1/2$. The ellipse is centered about the estimated values $\widehat{\boldsymbol{\theta}}$, and the tangents to the ellipse give the standard deviations of the estimators, σ_i and σ_j. The angle of the major axis of the ellipse is given by

$$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_j^2 - \sigma_i^2} \ , \tag{1.55}$$

where $\rho_{ij} = {\rm cov}[\widehat{\theta}_i, \widehat{\theta}_j]/\sigma_i\sigma_j$ is the correlation coefficient.

The correlation coefficient can be visualized as the fraction of the distance σ_i from the ellipse's horizontal centerline at which the ellipse becomes tangent to vertical, *i.e.*, at the distance $\rho_{ij}\sigma_i$ below the centerline as shown. As ρ_{ij} goes to $+1$ or -1, the ellipse thins to a diagonal line.

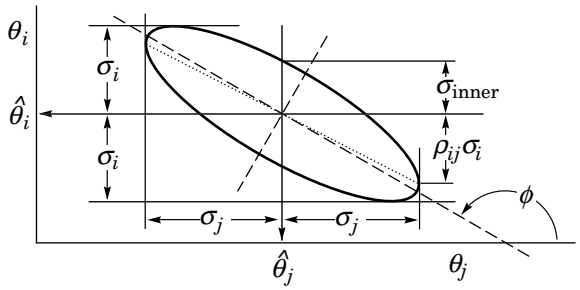

Figure 1.5: Standard error ellipse for the estimators $\widehat{\theta}_i$ and $\widehat{\theta}_j$. In this case the correlation is negative.

As in the single-variable case, because of the symmetry of the Gaussian function between $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$, one finds that contours of constant $\ln L$ or χ^2 cover the true values with a certain, fixed probability. That is, the confidence region is determined by

$$\ln L\,(\boldsymbol{\theta}) \geq \ln L_{\rm max} - \Delta \ln L \ , \tag{1.56}$$

or where a χ^2 has been defined for use with the method of least-squares,

$$\chi^2\,(\boldsymbol{\theta}) \leq \chi^2_{\rm min} + \Delta\chi^2 \ . \tag{1.57}$$

Values of $\Delta\chi^2$ or $2\Delta \ln L$ are given in Table 36.2 for several values of the coverage probability and number of fitted parameters.

For finite data samples, the probability for the regions determined by equations (1.56) or (1.57) to cover the true value of $\boldsymbol{\theta}$ will depend on $\boldsymbol{\theta}$, so these are not exact confidence regions according to our previous definition.

1.3.2.5. *Poisson or binomial data:* Another important class of measurements consists of counting a certain number of events, n. In this section, we will assume these are all events of the desired type, *i.e.*, there is no background. If n represents the number of events produced in a reaction with cross section σ, say, in a fixed integrated

Table 1.2: $\Delta\chi^2$ or $2\Delta \ln L$ corresponding to a coverage probability $1-\alpha$ in the large data sample limit, for joint estimation of m parameters.

$(1-\alpha)$ (%)	$m=1$	$m=2$	$m=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

luminosity $\mathcal{L}$, then it follows a Poisson distribution with mean $\nu = \sigma\mathcal{L}$. If, on the other hand, one has selected a larger sample of N events and found n of them to have a particular property, then n follows a binomial distribution where the parameter p gives the probability for the event to possess the property in question. This is appropriate, *e.g.*, for estimates of branching ratios or selection efficiencies based on a given total number of events.

For the case of Poisson distributed n, the upper and lower limits on the mean value ν can be found from the Neyman procedure to be

$$\nu_{\rm lo} = \tfrac{1}{2} F^{-1}_{\chi^2}\left(\alpha_{\rm lo}; 2n\right) , \tag{1.59a}$$

$$\nu_{\rm up} = \tfrac{1}{2} F^{-1}_{\chi^2}\left(1-\alpha_{\rm up}; 2\left(n+1\right)\right) , \tag{1.59b}$$

where the upper and lower limits are at confidence levels of $1-\alpha_{\rm lo}$ and $1-\alpha_{\rm up}$, respectively, and $F^{-1}_{\chi^2}$ is the *quantile* of the χ^2 distribution (inverse of the cumulative distribution). The quantiles $F^{-1}_{\chi^2}$ can be obtained from standard tables or from the CERNLIB routine `CHISIN`. For central confidence intervals at confidence level $1-\alpha$, set $\alpha_{\rm lo} = \alpha_{\rm up} = \alpha/2$.

It happens that the upper limit from Eq. (1.59*a*) coincides numerically with the Bayesian upper limit for a Poisson parameter, using a uniform prior p.d.f. for ν. Values for confidence levels of 90% and 95% are shown in Table 1.3. For the case of binomially distributed n successes out of N trials with probability of success p, the upper and lower limits on p are found to be

$$p_{\rm lo} = \frac{nF^{-1}_F\left[\alpha_{\rm lo}; 2n, 2\left(N-n+1\right)\right]}{N-n+1+nF^{-1}_F\left[\alpha_{\rm lo}; 2n, 2\left(N-n+1\right)\right]} , \tag{1.60a}$$

$$p_{\rm up} = \frac{\left(n+1\right)F^{-1}_F\left[1-\alpha_{\rm up}; 2\left(n+1\right), 2\left(N-n\right)\right]}{\left(N-n\right)+\left(n+1\right)F^{-1}_F\left[1-\alpha_{\rm up}; 2\left(n+1\right), 2\left(N-n\right)\right]} . \tag{1.60b}$$

Here F^{-1}_F is the quantile of the F distribution (also called the Fisher–Snedecor distribution; see [4]).

1.3.2.6. *Difficulties with intervals near a boundary:*

A number of issues arise in the construction and interpretation of confidence intervals when the parameter can only take on values in a restricted range. Important examples are where the mean of a Gaussian variable is constrained on physical grounds to be non-negative and where the experiment finds a Poisson-distributed number of events, n, which includes both signal and background. Application

Table 1.3: Lower and upper (one-sided) limits for the mean ν of a Poisson variable given n observed events in the absence of background, for confidence levels of 90% and 95%.

	$1-\alpha$ =90%		$1-\alpha$ =95%	
n	$\nu_{\rm lo}$	$\nu_{\rm up}$	$\nu_{\rm lo}$	$\nu_{\rm up}$
0	–	2.30	–	3.00
1	0.105	3.89	0.051	4.74
2	0.532	5.32	0.355	6.30
3	1.10	6.68	0.818	7.75
4	1.74	7.99	1.37	9.15
5	2.43	9.27	1.97	10.51
6	3.15	10.53	2.61	11.84
7	3.89	11.77	3.29	13.15
8	4.66	12.99	3.98	14.43
9	5.43	14.21	4.70	15.71
10	6.22	15.41	5.43	16.96

of some standard recipes can lead to intervals that are partially or entirely in the unphysical region. Furthermore, if the decision whether to report a one- or two-sided interval is based on the data, then the resulting intervals will not in general cover the parameter with the stated probability $1-\alpha$.

Several problems with such intervals are overcome by using the unified approach of Feldman and Cousins [31]. Properties of these intervals are described further in the *Review*. Table 36.4 gives the unified confidence intervals $[\nu_1, \nu_2]$ for the mean of a Poisson variable given n observed events in the absence of background, for confidence levels of 90% and 95%. The values of $1-\alpha$ given here refer to the coverage of the true parameter by the whole interval $[\nu_1, \nu_2]$. In Table 1.3 for the one-sided upper and lower limits, however, $1-\alpha$ referred to the probability to have individually $\nu_{\rm up} \geq \nu$ or $\nu_{\rm lo} \leq \nu$.

Another possibility is to construct a Bayesian interval as described in Section 1.3.1. The presence of the boundary can be incorporated simply by setting the prior density to zero in the unphysical region. Advantages and pitfalls of this approach are discussed further in the *Review*.

Another alternative is presented by the intervals found from the likelihood function or χ^2 using the prescription of Equations (1.56) or (1.57). As in the case of the Bayesian intervals, the coverage probability is not, in general, independent of the true parameter. Furthermore, these intervals can for some parameter values undercover.

In any case it is important to report sufficient information so that the result can be combined with other measurements. Often this means giving an unbiased estimator and its standard deviation, even if the estimated value is in the unphysical region. It is also useful to report the likelihood function or an appropriate summary of it. Although this by itself is not sufficient to construct a frequentist

Table 1.4: Unified confidence intervals $[\nu_1, \nu_2]$ for a the mean of a Poisson variable given n observed events in the absence of background, for confidence levels of 90% and 95%.

	$1-\alpha$ =90%		$1-\alpha$ =95%	
n	ν_1	ν_2	ν_1	ν_2
0	0.00	2.44	0.00	3.09
1	0.11	4.36	0.05	5.14
2	0.53	5.91	0.36	6.72
3	1.10	7.42	0.82	8.25
4	1.47	8.60	1.37	9.76
5	1.84	9.99	1.84	11.26
6	2.21	11.47	2.21	12.75
7	3.56	12.53	2.58	13.81
8	3.96	13.99	2.94	15.29
9	4.36	15.30	4.36	16.77
10	5.50	16.50	4.75	17.82

confidence interval, it can be used to find the Bayesian posterior probability density for any desired prior p.d.f.

Further discussion and all references may be found in the full *Review of Particle Physics;* the equation and reference numbering corresponds to that version.